Overview of the TREC 2014 Session Track

Ben Carterette* Evangelos Kanoulas† Mark Hall‡ Paul Clough§

1 Introduction

The TREC Session track ran for the fourth time in 2014. The track has the primary goal of providing test collections and evaluation measures for studying information retrieval over user *sessions* rather than one-time queries. These test collections are meant to be portable, reusable, statistically powerful, and open to anyone that wishes to work on the problem of retrieval over sessions.

The experimental design of the track was similar to that of the previous three years [5, 6, 1]:

- sessions were real user sessions with a search engine that include queries, retrieved results, clicks, and dwell times;
- retrieval tasks were designed to study the effect of using session data in retrieval for only the mth query in a session.

For the 2014 track, sessions were obtained from workers on Amazon's Mechanical Turk. As a result, the 2014 data includes far more sessions than previous years—1,257 unique sessions as compared to around 100 for each of the previous three years. Apart from that, there is little different from the 2013 track [1].

This overview is organized as follows: in Section 2 we describe the tasks participants were to perform. In Section 3 we describe the corpus, topics, and sessions that comprise the test collection. Section 4 gives some information about submitted runs. In Section 5 we describe relevance judging and evaluation measures, and Sections 6 present evaluation results and analysis.

2 Evaluation Tasks

We use the word "session" to mean a sequence of reformulations along with any user interaction with the retrieved results in service of satisfying a specific topical information need. The primary goal for participants of the 2013 track was to provide the best possible results for the last query in a session given data from the user interactions leading up to it.

*Department of Computer & Information Sciences, University of Delaware, Newark, DE, USA
†Google, Zurich, Switzerland
‡Department of Computing, Edge Hill University, Ormskirk, UK
§Information School, University of Sheffield, Sheffield, UK

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
NOV 2014		**00-00-2014 to 00-00-2014**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Overview of the TREC 2014 Session Track	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **University of Delaware,Department of Computer & Information Sciences,Newark,DE,19716**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
presented in the proceedings of the Twenty-Third Text REtrieval Conference (TREC 2014) held in Gaithersburg, Maryland, November 19-21, 2014. The conference was co-sponsored by the National Institute of Standards and Technology (NIST) and the Defense Advanced Research Projects Agency (DARPA).

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**15**	

We provided participants with a set of 1,257 sessions of varying length (described in more detail in Section 3). 1,021 of these were targeted for evaluation; the remaining 236 could be used for training, as could any data from the previous two years of the track. Each of the 1,021 evaluated sessions, numbered $i = 1..1021$, consists of:

- m_i blocks of user interactions (the *length* of the session);
- the final query in the session q_{m_i}, referred to as the *current query*;
- $m_i - 1$ blocks of user interactions in the session prior to the current query:
  1. the set of past queries in the session, q_1, q_2, ..., q_{m_i-1};
  2. the ranked list of URLs seen by the user for each of those queries;
  3. the set of clicked URLs/snippets.

In addition, each user action is accompanied by a time relative to the start of the session. The remaining 236 sessions consist of the same data except all of them have length $m_i = 1$ and thus they have no current/final query.

Participants were to run the 1,021 current queries against their search engines under each of the following three conditions separately:

RL1 ignoring the session prior to this query

RL2 considering all the items (1), (2) and (3) above for the current session, i.e the queries prior to the current, the ranked lists of URLs and the corresponding web pages, the clicked URLs and the time spent on any interaction

RL3 considering all data in the entire session log (in particular, other user sessions on the same topic)

Comparing the retrieval effectiveness in (RL1) with the retrieval effectiveness in (RL2)–(RL3), one can evaluate the effectiveness of different algorithms for incorporating session information into retrieval.

3 Test Collection

Like most IR test collections, ours consists of a corpus, a set of topics, and relevance judgments (described in the next section). Unlike most test collections, ours also includes a set of *sessions* of user interactions (including query reformulations). A single topic can have more than one session associated with it, since two different users could go about satisfying the same information need in very different ways and with different degrees of success.

3.1 Corpus

The track used the ClueWeb12 collection. The full collection consists of roughly 730 million English-language web pages, comprising approximately 5TB of compressed data. The dataset was crawled from the Web during February and March 2012.

Participants were encouraged to use the entire collection, however submissions over the smaller "Category B" collection of about 35 million documents were accepted. Note that Category B submissions was evaluated as if they were Category A submissions.

3.2 Topics

For the 2014 track, we re-used topics that had been defined for the 2012 and 2013 tracks. Both of these topic sets were defined with an attempt to control two facets of search tasks as defined by Li and Belkin [7]: "product" and "goal quality". The "product" facet represents the end goal of the task, which is either *intellectual*–to produce new ideas or findings (e.g. learn about a topic or make decisions based on information collected)–or *factual*–locating facts, data, or other information items. An example of such variation (from the 2012 Session track topics) can be viewed below.

Query: dehumidifiers

- Session topic: 35
  "Product" facet: Factual
  Description: You would like to buy a dehumidifier. What are some of the technical specifications you should be looking at? What is the price range for dehumidifiers? What makes one dehumidifier more expensive than another?

- Session topic: 37
  "Product" facet: Intellectual
  Description: You would like to buy a dehumidifier. On what basis should you compare different dehumidifiers?

The "goal quality" facet reflects *specific goal(s)* and *amorphous goal(s)*. This is very similar to the dimension that Ingwersen and Järvelin [4] proposed as well-defined and ill-defined information need. Tasks with specific goals have a well-defined information need, while in tasks with amorphous goals, the information need is ill-defined. Tasks with amorphous goals might require users to redefine the topic or identify specific aspects of the subject themselves. An example (again from the 2012 track) can be viewed below:

Query: Swahili dishes

- Session topic: 38
  "Goal" facet: specific goals
  Description: What are some traditional Swahili dishes? What ingredients do they use to cook them? Are swahili people using any particular herb in their dishes? Could you find these ingredients in your country? Are there any recipes you can find online?

- Session topic: 40
  "Goal" facet: amorphous goals
  Description: One of your friends from Kenya invited you to attend a party in his house and have a taste of traditional swahili dishes. You would like to search and find some information about Swahili dishes.

Combining "product" and "goal quality" facets generates four classes of topics, characterized as follows: a factual task with specific goals is *known-item search*; a factual task with amorphous goals is *known-subject search*; an intellectual task with specific goals is *interpretive search*; and an intellectual task with amorphous goals is *exploratory search*. A complete example of all four combinations can be viewed below:

Query: depression symptoms

- Task: Known-item search
  Description: What is depression? What are the major symptoms of depression? What medications, therapies and other treatments can be used to treat depression symptoms? Who performs therapy and what are the costs? Does health insurance pay for any of the treatments?
- Task: Known-subject search
  Description: You think that one of your friends may have depression, and you want to search information about the depression symptoms and possible treatments.
- Task: Interpretive search
  Description: Depression is a loaded word in our culture. What are the symptoms that could differentiate depression from having just a bad month of excessive emotions? When should one seek help and what kind?
- Task: Exploratory search
  Description: A friend has been complaining for months that she is unhappy with her life. She has also mentioned that she can't easily sleep at nights. You think that she may be suffering from depression. You want to understand if this is the case and how you could assist her in getting some help from medical professionals.

For the 2014 track, we selected topics from 2012 and 2013 such that there would be the same number of topics in each of the four categories. We biased selection to topics that had longer user sessions and more clicks, i.e. more user interaction overall.

3.3 Sessions

As describe above, a session is a series of actions, including queries and clicks on ranked results, that a user performs in the process of trying to satisfy the information need represented by the topic. The topics were presented to actual users, who were then able to use a fully-functional custom search engine for ClueWeb12 in order to satisfy the information need described by the topic. .

The search system used an indri index of ClueWeb12 as a backend. Each of the 20 ClueWeb12 segments (ClueWeb12_00 through ClueWeb12_19) was indexed using the Krovetz stemmer and no stopword list. The indexes searched contained only text from title fields, anchor text from incoming links ("inlink" text), and page URLs. We chose to index these fields only in an effort to get faster response times.

Each query was plugged into an indri query language template as exemplified below for the query "quitting smoking":

```
#combine(#weight(1   #combine(quitting smoking)
                 1   #weight(1 #combine(quitting.title smoking.title)
                               1 #uw(quitting smoking).title)
                 50  #weight(1 #combine(quitting.inlink smoking.inlink)
                               1 #uw(quitting smoking).inlink)
                 0.5 #weight(1 #combine(quitting.url smoking.url)
                               1 #uw(quitting smoking).url)))
```

The retrieval score is computed from four component models: a basic query-likelihood model for the full document representation and three weighted combinations of basic query-likelihood field models with unordered-window within-field models. The "inlink" model was weighted 50 times higher than the title model, and 100 times higher than the URL model. This query template is the product of manual search and investigation of retrieved results.

The search interface connected to our indri backend to retrieve the top 50 results (which were filtered further so that at most two documents from any domain would be shown). Users were shown these results in pages of 10 along with a snippet produced by indri's built-in snippet generator. They could click results to see the current version of the page, which of course could be different from the version in the index. We asked users to search for a minimum of 3 minutes.

Our full indri search engine can be quite slow, requiring 30 seconds or more to respond to some queries. We cached results for all queries entered and served results from the cache when available. In order to ensure a relatively fast response for all queries, we also built a smaller index consisting of documents in cached results. Then, if no cached results were available, the query was submitted to both indexes, which were given a maximum of 6 seconds to respond. If neither had responded after 6 seconds, results were assembled in an ad hoc way by fusing cached results from previous queries that used the same terms and phrases.

The system recorded the user's interactions with the retrieval system, including the queries issued, query reformulations, and items clicked in the results page. When data collection was complete, we had acquired a set of candidate sessions to go with the candidate topics we defined above. Each session consists of a topic, a set of queries actual users posed about the topic, the retrieved results, and the user interactions with the retrieved results.

Session data is provided in an XML file. Part of an example session is shown on page 6.

The released data comprised 1,257 full sessions. 1,021 of these have at least one reformulation (i.e. at least two queries, of which the second is the `currentquery`); these were the ones targeted for evaluation. Of these, 185 have at least three reformulations, 145 have at least four, 107 have at least five, 17 have at least 10, and there is one session with 15 reformulations. On average there are 4.33 queries per session, and the median session length is 2 queries. These are substantially shorter sessions than 2013, most likely reflecting the different user base. The median length of time spent on a session was about 2.8. minutes. The data includes 1,685 total clicks across 1,257 sessions, 1.34 per session on average and about one click every other user query.

```
<session num="10" starttime="0">
   <topic num="12">
      <desc>Your friend would like to quit smoking. You would like to provIde him with
            relevant information about: the different ways to quit smoking, programs
            available to help quit smoking, benefits of quitting smoking, second effects
            of quitting smoking, using hypnosis to quit smoking, using the cold turkey
            method to quit smoking</desc>
   </topic>
   <interaction num="1" starttime="8.30123">
      <query>quit smoking</query>
      <results>
         <result rank="1">
            <url>http://quitsmoking.about.com</url>
            <clueweb12id>clueweb12-0005wb-77-27713</clueweb12id>
            <title>Quit Smoking | Quit Smoking Support | Smoking Cessation</title>
            <snippet>Quit Smoking | Quit Smoking Support | Smoking Cessation
            About.comHealthSmoking Cessation Smoking Cessation Search Smoking
            CessationHealth RisksHow to QuitAfter Quitting Share Quit Smoking
            ToolboxShocking Tobacco Facts10 Things to Avoid When You Quit Guide since
            2003Terry MartinSmoking Cessation GuideSign up for My NewsletterMy Bio...</snippet>
         </result>
         ...
         <result rank="10">
            <url>http://www.heart.org/HEARTORG/GettingHealthy/QuitSmoking/Quit-Smoking_UCM_001085_Su
            <clueweb12id>clueweb12-0300tw-20-20611</clueweb12id>
            <title>Quit Smoking</title>
            <snippet>Quit Smoking ...0 Grams Trans Fat Oils and Fats Restaurant FAQs
            Other Restaurant Resources Quit Smoking Smoking Smoking Life quit today. We
            can help. Learn more. Quit Smoking?Smoking is the most important preventable c
            ause of premature death in the United...</snippet>
         </result>
      </results>
      <clicked>
         <click num="1" starttime="12.984659" endtime="20.557844">
            <rank>3</rank>
         </click>
         <click num="2" starttime="27.030967" endtime="55.220869">
            <rank>1</rank>
         </click>
         <click num="3" starttime="55.220869" endtime="60.704926">
            <rank>6</rank>
         </click>
         <click num="4" starttime="60.704926" endtime="69.165489">
            <rank>5</rank>
         </click>
      </clicked>
   </interaction>
   <currentquery starttime="78.226578">
      <query>quit smoking cold turkey</query>
   </currentquery>
</session>
```

4 Submissions

Participating sites were permitted to submit up to three runs. Each submitted run includes up to three separate ranked result lists for all 1,022 sessions. Files were named "runTag.RLn", where "runTag" is a unique identifier for the site and the particular submission, and "RLn" is RL1, RL2, or RL3, depending on the experimental condition.

The track received 27 runs from 11 groups, for a total of 74 ranked lists for each session. They are listed in Table 1.

num	group	group code
1.	Bauhaus-Universitat Weimar, Germany	BUW
2.	Endicott College, USA	Endicott
3.	Georgetown University, USA	Georgetown
4.	Institute of Computing Technology, Chinese Academy of Sciences, China	ICTNET
5.	University of Delaware, USA	udel
6.	University of Massachusetts Amherst, USA	UMASS_CIIR
7.	Siena College, USA	SCIAITeam
8.	East China Normal University, China	ecnu
9.	Tianjin Univesity, China	wlzn_tju_cn
10.	University College London, UK	UCL
11.	Leidos, USA	RAMA

Table 1: Groups participating in the 2014 Sessions Track.

5 Session Evaluation

5.1 Relevance Judgments

Judging was done by assessors at NIST. As described above, each topic was the subject of one or more sessions. Thus pools for judging were formed by topic, not by session.

For each topic, a pool was formed from the ranked results produced by our indri system for the past queries $q_1 ... q_{m-1}$ and the current query q_m, along with the top 10 ranked documents from the submitted runs on the current query q_m.Because there was such a large number of sessions, only the first 100 (by ID) contributed to the pool. These sessions include 51 of the 60 topics, so there are 9 that received no relevance judgments this year.

The NIST assessors judged each document in the pool with respect to the topic description. URLs were sorted by domain prior to judging so that assessors would see all pages from the same domain before moving to another one.

The qrels produced have the following format:

```
<topic-id> 0 <doc-id> <judgment>
```

Judgment values are: -2 for spam document (i.e. the page does not appear to be useful for any reasonable purpose; it may be spam or junk.); 0 for not relevant (i.e. the content of this page does not provide useful information on the topic, but may provide useful information on other topics, including other interpretations of the same query); 1 for relevant (i.e. the content of this page provides some information on the topic, which may be minimal; the relevant information must be on that page, not just promising-looking anchor text pointing to a possibly useful page); 2 for highly relevant (i.e. the content of this page provides substantial information on the topic); 3 for key, (i.e. the page or site is dedicated to the topic; authoritative and comprehensive, worthy of being a top result in a web search engine; typically, key pages are more comprehensive, have higher quality, and are from more trustworthy sources than the merely highly relevant page); and 4 for navigational (i.e. this page represents a home page of an entity directly named by the query; the user may be searching for this specific page or site; there is often at most one page that deserves a Navigational judgment for an aspect).

Relevance judgments were eventually transformed to relevance grades with spam and non-relevant documents assigned a grade of 0, relevant assigned a grade of 1, highly relevant assigned a grade of 2, key assigned a grade of 3, and navigational assigned a grade of 4.

A total of 16,949 pages were judged. Out of these 16,949 pages, 31 were judged as navigational, 148 as key, 1,160 as highly relevant, 3,536 as relevant, 11,277 as nonrelevant, and 797 as spam. On average there were 121 nonrelevant (or spam) and 49 relevant (across all types of relevance) documents per session. Only 5 of the topics had at least one "key" document, and 24 had at least one "nav" document.[1]

5.2 Evaluation Measures

Based on the qrels provided by NIST and the decisions described above, we evaluated submitted runs by eight measures:

- Expected Reciprocal Rank (ERR) [3]
- ERR@10
- ERR normalized by the maximum ERR per query (nERR)
- nERR@10
- nDCG
- nDCG@10
- Average Precision (AP)
- Precision@10

run	RL1	RL2		RL3	
udel14Run1	0.2262	–		0.2432	↑
GUS14Run3	0.2053	0.2482	⇑	0.2580	↑
GUS14Run2	0.2053	0.2482	⇑	0.2458	↓
GUS14Run1	0.2053	0.2458	⇑	0.2443	↓
ICTNET14SER1	0.1890	0.2357	⇑	0.2431	↑
ICTNET14SER3	0.1890	0.2288	⇑	0.2356	↑
ecnusession1	0.1890	0.2091	↑	0.2157	↑
webisSt14ax	0.1890	0.2025	↑	0.1985	↓
ICTNET14SER2	0.1890	0.1976	↑	0.2045	↑
webis2014act	0.1890	0.1886	↓	0.2040	↑
webis2014db	0.1890	0.1876	↓	0.2067	⇑
ECxCGxPRF	0.1810	0.1815	↑	0.1932	↑
ECxSRMxOS	0.1791	0.1951	↑	0.2047	↑
ECxSRMxPRF	0.1683	0.1784	↑	0.1939	↑
SCIAITeamC	0.1650	0.1701	↑	–	
SCIAITeamL	0.1650	0.1690	↑	–	
SCIAITeamF	0.1650	0.1137	⇓	–	
UMASS3	0.1630	0.1832	↑	0.1832	↔
UMASS1	0.1630	0.1714	↑	0.1702	↓
UMASS2	0.1630	0.1496	↓	0.1349	↓
UMASS4	0.1630	0.1353	↓	0.1414	↑
uclbaseline	0.1580	–		–	
udelitu	0.1515	–		–	
ecnusession2	0.1390	0.2139	⇑	0.2163	↑
WLZNTJU	0.1365	0.1467	↑	0.1453	↓
RAMA	0.1342	0.1327	↓	0.0917	↓
ecnusession3	0.0815	0.1996	⇑	0.2322	⇑

Table 2: All results by nDCG@10 for the current query in the first 100 sessions for each condition (sorted in decreasing order of RL1 nDCG@10). Boldface indicates the highest nDCG@10 in the condition. $\uparrow, \downarrow$ indicate positive or negative differences from the prior condition. $\Uparrow, \Downarrow$ indicate statistically significant ($p < 0.05$ by a paired two-sided t-test) positive or negative differences from the prior condition. $\leftrightarrow$ indicates no difference from the prior condition.

6 Evaluation Results

Table 2 shows all results (by nDCG@10) for all submitted runs in all three experimental conditions. If RL1 (no information about the session) is the baseline, most (18 of 24 that submitted an RL2) of the submitted runs were able to improve on that using information about the session prior to the last query (RL2 results). Seven of those were statistically significant improvements. A smaller set of systems (14 of 22 that submitted an RL3) were further able to improve by making use of the full log (RL3 results), though only two of these were statistically significant over RL2.

[1]Somewhat strangely, two topics had more than 10 "key" documents, and one had more than 35 "nav" documents.

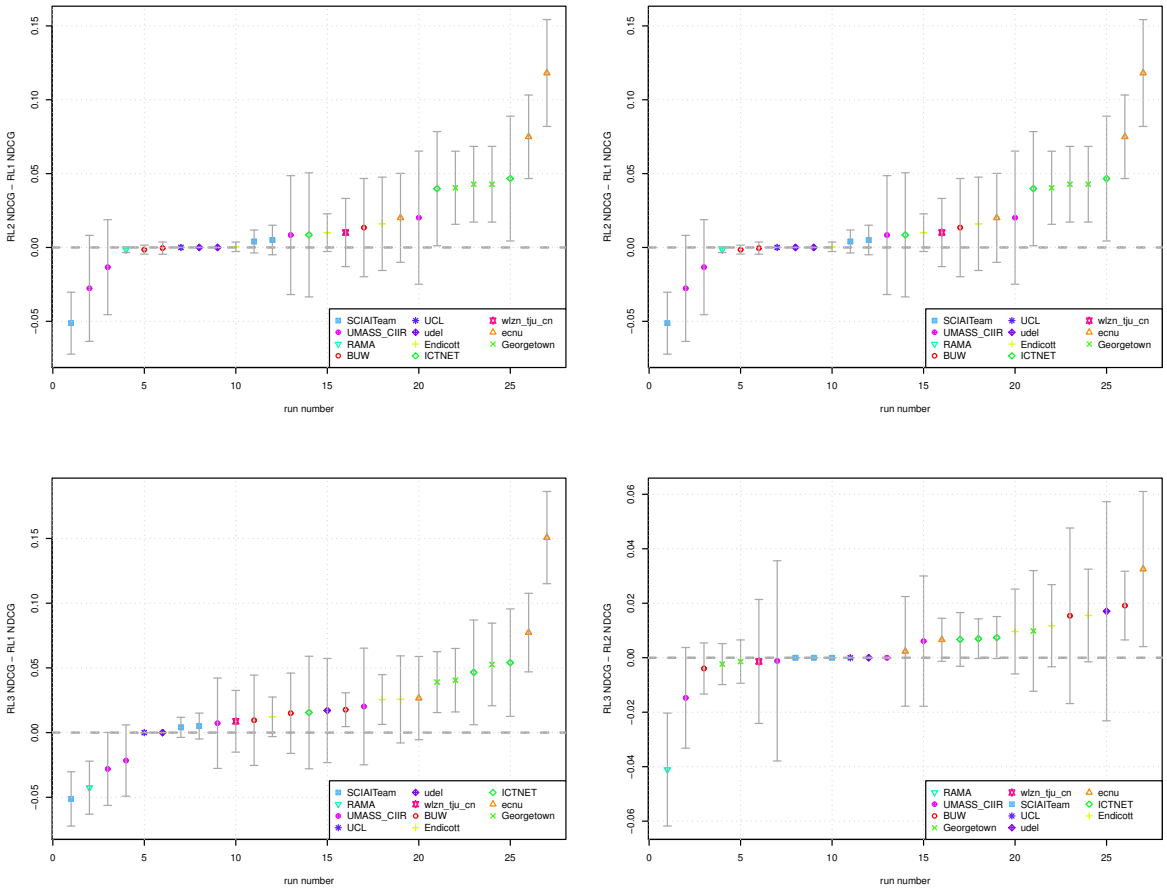

Figure 1: Left: Changes in nDCG@10 from RL1 to (from top to bottom) RL2 and RL3. Right: Changes in nDCG@10 from RL1 to RL2 and RL2 to RL3. Error bars are 95% confidence intervals.

Figure 1 shows changes in nDCG@10 from the RL1 baseline (left) or with increasing information (right). The plots going down the left column show changes in nDCG@10 from using no previous data (RL1) to using greater and greater amounts of previous data. The dashed line is a difference in nDCG of zero; points above that line represent systems that saw an improvement from using the additional data while points below it represent systems that were hurt with the additional data. The 95% confidence intervals give a rough idea of whether the results are significant.

On the right-hand side, Figure 1 shows changes in nDCG@10 with increasing amounts of previous data: going from RL1 to RL2, then RL2 to RL3. Only a few systems see improvement at every step, and the improvements are not significant.

Figure 2 shows the same information for unnormalized ERR. The two measures are well-correlated, with Kendall's τ correlation of 0.83 between the nDCG@10 difference from RL1 to RL2 and the ERR difference from RL1 to RL2.

Table 3 shows results for all sessions that each run ranked documents for. Some groups chose to return results for only the first 240 sessions, while others returned results for all 1021 sessions—the

run	sessions	RL1	RL2		RL3	
GUS14Run3	1021	0.1539	0.1917	↑	0.2099	↑
GUS14Run2	1021	0.1539	0.1917	↑	0.1952	↑
GUS14Run1	1021	0.1539	0.1911	↑	0.1949	↑
webis2014act	240	0.1349	0.1350	↑	0.1349	↓
ecnusession1	240	0.1328	0.1638	↑	0.1665	↑
SCIAITeamC	240	0.1210	0.1239	↑	–	
SCIAITeamL	240	0.1210	0.1234	↑	–	
SCIAITeamF	240	0.1210	0.0808	↓	–	
udel14Run1	1021	0.1195	–		0.1947	
webis2014db	240	0.1164	0.1340	↑	0.1565	↑
ICTNET14SER1	1021	0.1164	0.1232	↑	0.1284	↑
ICTNET14SER3	1021	0.1164	0.1228	↑	0.1277	↑
ICTNET14SER2	1021	0.1164	0.1135	↓	0.1137	↑
webisSt14ax	1021	0.1164	0.0955	↓	0.0936	↓
ECxCGxPRF	1021	0.1054	0.1062	↑	0.1123	↑
ECxSRMxOS	1021	0.1049	0.0911	↓	0.1011	↑
UMASS3	240	0.1034	0.1185	↑	0.1185	↔
UMASS1	240	0.1034	0.1170	↑	0.1229	↑
UMASS2	240	0.1034	0.0935	↓	0.0854	↓
UMASS4	240	0.1034	0.0840	↓	0.0936	↑
ecnusession2	240	0.0999	0.1656	↑	0.1698	↑
uclbaseline	240	0.0981	–		–	
udelitu	1021	0.0972	–		–	
WLZNTJU	240	0.0892	0.0912	↑	0.0987	↑
ECxSRMxPRF	1021	0.0847	0.0792	↓	0.0911	↑
RAMA	1021	0.0798	0.0793	↓	0.0620	↓
ecnusession3	240	0.0203	0.1422	↑	0.1697	↑

Table 3: All results by nDCG@10 for the current query in all ranked sessions for each condition (sorted in decreasing order of RL1 nDCG@10).

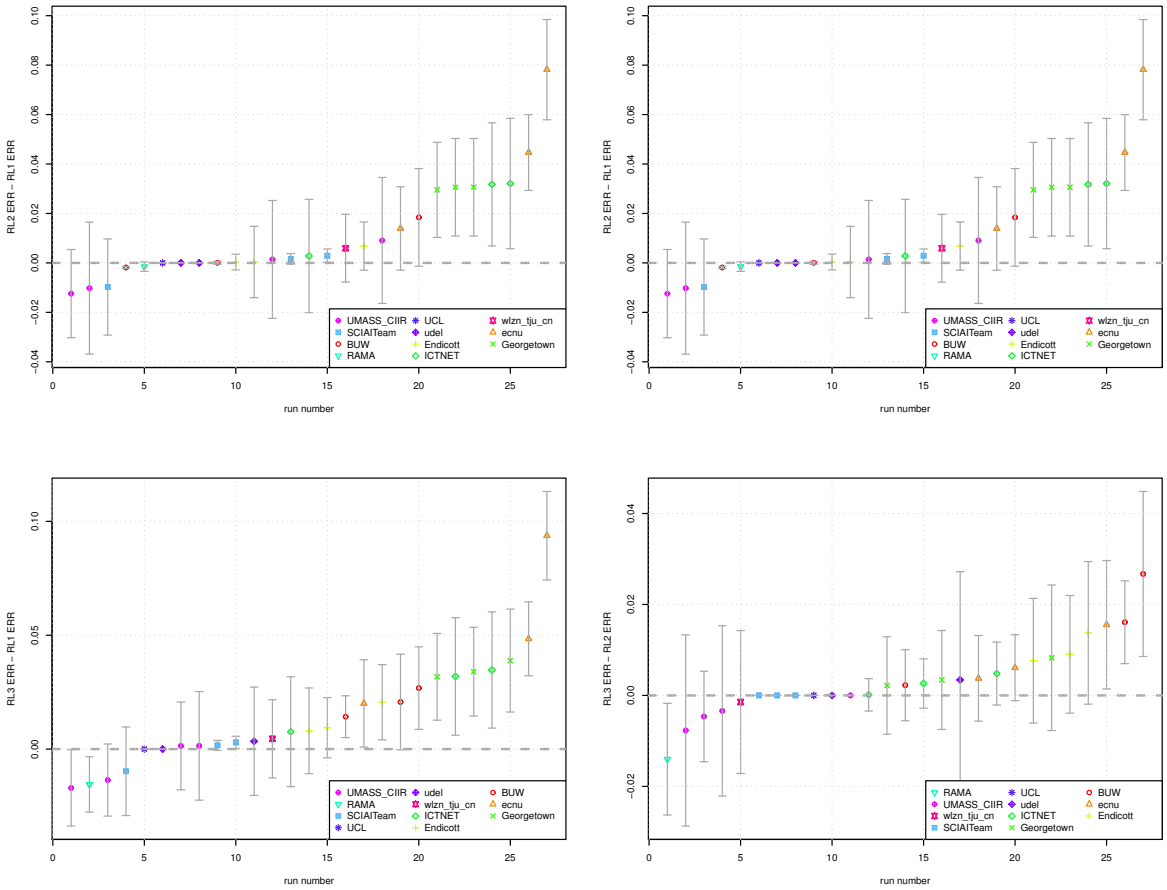

Figure 2: Left: Changes in ERR from RL1 to (from top to bottom) RL2 and RL3. Right: Changes in ERR from RL1 to RL2 and RL2 to RL3. Error bars are 95% confidence intervals.

split is about even. Note that since only the first 100 sessions were used to form pools, evaluation measures in this table have many missing judgments. We have not yet analyzed the extent to which that affects results. However, we observe that the Kendall's τ rank correlation between the RL1 evaluation in Table 2 and Table 3 is only 0.69, suggesting that the missing judgments might make a big difference.

7 Analysis

1. By topic category—effectiveness, other stuff
2. By user base
3. By underlying search engine

	2011	2012	2013	2014
collection	ClueWeb09	ClueWeb09	ClueWeb12	ClueWeb12
topic properties				
topic set size				
topic cat. dist.	known-item			
session properties				
user population	U. Sheffield	U. Sheffield	U. Sheffield + IR researchers	Mechanical Turk
search engine	BOSS	BOSS	indri	indri
sessions				
sessions per topic				
mean/median length				
median time				
mean num clicks				
relevance judgments				
topics judged				
sessions in pool				
total judgments				
evaluation by nDCG@10				
max RL[34] - RL1				

7.1 Topic reuse

As mentioned in Section **??**, topics for 2014 were taken from the 2012 and 2013 Session tracks. Since 2013 also used ClueWeb12, we can compare 2013 and 2014 data for these topics.

7.1.1 User queries

We looked at similarity between user queries from 2013 to 2014. First, for each topic we estimate two unigram language models: one for queries issued for 2013 topics and one for queries issued for 2014 topics. Model parameters are estimated as follows:

Examples of topic language models are shown in Table **??**. The two examples show two extreme cases: high agreement about query terms (topic 14) and low agreement about query terms (topic 32). Topic 32 in particular is an example of how difference in prior knowledge can influence search behavior: the topic asks for good places in the US to travel with a lot to do in a 150-mile radius. 2013 users immediately thought of specific cities in densely-populated areas like Philadelpha and New York City, while 2014 users use the much more vague phrase "best cities in united states".

On average, the difference in topic models is greatest for "exploratory" topics, followed by "interpretive" topics, followed by "known-subject" topics, followed by "known-item" topics (though these differences are not significant, as the sample size is small). This confirms our intuition.

topic 14				topic 32			
term	$P_{2013}(w\|T)$	term	$P_{2014}(w\|T)$	term	$P_{2013}(w\|T)$	term	$P_{2014}(w\|T)$
naturalism	0.250	god	0.183	maps	0.078	states	0.096
god	0.188	naturalism	0.111	tourism	0.078	united	0.096
theory	0.125	of	0.085	philadelphia	0.078	in	0.067
existence	0.125	theory	0.078	travel	0.065	best	0.067
not	0.063	evidence	0.059	nyc	0.065	cities	0.059

2013 \ 2014	rel	not
relevant	272	84
not relevant	220	357

7.1.2 Retrieved documents

7.1.3 Assessor agreement

Since topics are the same, and we made no effort to remove documents judged for 2013 from the pools for 2014, it is likely that documents judged for 2013 were re-judged for 2014. This allows us to calculate agreement between assessors.

27 of the 60 topics for 2014 were taken from the 2013 data. Of these 27, 24 have relevance judgments.

In 2013, a total of 8,125 documents were judged for those 27 topics. In 2014, a total of 7,894 documents were judged for the 24 matching topics. That is a total of 16,019 documents judged. Of these, only 933 were judged both years. That is, *only 6% of documents judged for these topics* were judged both years. This is remarkably low and speaks poorly to the reusability of the collections (see below).

Table 7.1.3 shows agreement on binary relevance (converting all judgments ≥ 1 to 1 and all judgmend ≤ 0 to 0) among these 933. Overall agreement is 67%. Agreement on documents that at least one assessor judged relevant is 47%, which is in line with previous studies on disagreement. Interestingly, assessors were much more likely to say a document judged nonrelevant in 2013 was relevant in 2014 than vice versa.

Table 7.1.3 shows agreement on relevance grades.

2013 \ 2014	4	3	2	1	0	-2
nav - 4	1	0	0	0	0	0
key - 3	0	1	2	7	4	0
hi - 2	0	4	28	52	14	2
rel - 1	1	12	75	89	64	0
not - 0	4	5	50	161	337	11
junk -2	0	0	0	0	4	5

References

[1] B. Carterette, E. Kanoulas, A. Bah, M. Hall, and P. D. Clough. Overview of the trec 2013 session track. In *Proceedings of TREC*, 2013.

[2] B. Carterette, E. Kanoulas, P. D. Clough, and M. Sanderson, editors. *Proceedings of the ECIR 2011 Workshop on Information Retrieval Over Query Sessions*, Available at `http://ir.cis.udel.edu/ECIR11Sessions`.

[3] O. Chapelle, D. Metzler, Y. Zhang, and P. Grinspan. Expected reciprocal rank for graded relevance. In *In Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM)*, 2009.

[4] P. Ingwersen and K. Järvelin. *The Turn: Integration of Information Seeking and Retrieval in Context (The Information Retrieval Series)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2005.

[5] E. Kanoulas, B. Carterette, M. Hall, P. Clough, and M. Sanderson. Session track 2011 overview. In *The Twentieth Text REtrieval Conference Proceedings (TREC 2011)*. National Institute of Standards and Technology, 2012. (`http://trec.nist.gov/pubs/trec20/papers/SESSION.OVERVIEW.2011.pdf`).

[6] E. Kanoulas, B. Carterette, M. Hall, P. D. Clough, and M. Sanderson. Overview of the trec 2012 session track. In *Proceedings of TREC*, 2012.

[7] Y. Li and N. J. Belkin. A faceted approach to conceptualizing tasks in information seeking. *Inf. Process. Manage.*, 44(6):1822–1837, Nov. 2008.